A New Tool for Automated Data Extraction

John D. Halamka M.D., Roger J. Lewis M.D.,Ph.D.
Harbor/UCLA Medical Center Torrance, California

Although many Hospital Information Systems (HIS) capture clinical laboratory results to facilitate clinical care, they do not provide the tools to organize and analyze the data for clinical trials, outcomes research and continuous quality improvement activities. Most such systems can export data as ASCII text, however. Realizing that the HIS data could be exported to a text file, we created the Automated Patient Information Extraction System (APEX) to link the exported HIS text with our clinical databases and analysis tools.

APEX is designed in Visual Basic with a standard Microsoft Windows graphical user interface. The system provides a set of powerful tools to selectively extract information from exported HIS text data and create structured query language (SQL) databases. APEX flexibly processes any text data which contain a patient identifier, a test identifier and a result. The order of these fields and the structure of the text file may be customized by the user. For example, our text data consisted of patient identifier, test name and test result in a quote/comma delimited file i.e. "535-23-1234","GLUCOSE", 250. The system provides two processing modes: dataset completion and new database creation.

The dataset completion function automatically adds laboratory and demographic data to research databases containing pre-enrolled patients.

APEX searches the exported HIS text and identifies specific patients by matching account numbers. A flexible graphical user interface enables mapping of HIS text data structures to research database structures. Numerous options are available such as selection of the first occurrence of each test result, the last occurrence of each test result, or interactive viewing and selection of test results.

The second APEX function is new database creation. This mode automatically enters patients into research databases based on clinical information found in the HIS data. In new database creation mode, criteria may be entered which qualify a patient for study inclusion. A full range of comparisons can be made on any HIS field such as AGE>5 AND AGE<75, GLUCOSE>500 OR GLUCOSE<100, and BICARB<=15.

Three additional tools complement APEX operations. A powerful ICD-9 tool enables new database creation based on ICD-9 diagnosis. Up to five keywords may be entered for search against the complete ICD-9 database. The 15,000 available 5 digit ICD-9 codes are searched and all codes contain these keywords are displayed for user review and editing. Once approved, these ICD-9 codes are automatically processed into selection criteria for the new database creation function. For example, entry of the keyword INFARCT generates 20 ICD-9 codes. With a single keystroke, the HIS text data can be automatically searched and all patients with these ICD-9 codes entered into the research database.

A SQL tool is also available in APEX. This tool allows further refinement of the research database though the use of standard database manipulation commands. Using SQL, multiple inclusion and exclusion criteria may be further refined and applied to research data.

Finally, an HIS file tool is available to prescan exported HIS text files and automatically determine their structure. The user specifies basic HIS text data layout information such as the characters used for field and record delimiters. APEX analyzes the HIS text file and displays sample processed data for user review. The user may modify all aspects of APEX data processing, making APEX flexible enough to work with any ASCII dataset from any HIS system.

As a specific test of the system we used APEX to populate a clinical research database. The particular study contained over 300 pre-enrolled patients and required capture of 102 variables per patient including laboratory data, process of care data and demographic data. Our HIS system stores results in a non-standard data structure not accessible by standard database manipulation tools. We exported 700 megabytes of text data from the Harbor-UCLA HIS system into quote and comma delimited text files containing three fields per record - patient identifier, test name and test result. The APEX system examined 400 days of hospital data representing 200,000 patient visits and completed our study database in 6 hours. A similar human effort would have taken hundreds of hours.

The APEX system provides a versatile method for adapting available HIS data to clinical research use. It can be used to complete a dataset, create a new database or to verify an existing dataset.

0195—4210/96/$5.00